Efficient Sharpe Ratio Tests: Exploiting the Precision of Volatility Estimates

March 16, 2026

Abstract

The Sharpe ratio combines expected return and volatility, two components that are learned at different statistical rates. Expected returns are learned slowly with calendar time, while volatility is typically estimated much more precisely from many return observations. We show how to remove or reduce the uncertainty that volatility estimates contribute to standard errors for Sharpe ratios.

Testing whether a Sharpe ratio equals zero is logically a test of the mean return only. Treating it as a ratio test injects unnecessary volatility noise and reduces statistical power.

For comparisons of positive Sharpe ratios and tests against performance thresholds, we use a delta-method framework that separates mean and volatility uncertainty, explicitly incorporates sampling frequency, and yields analytical results in special cases. Conventional analytical formulas do not incorporate sampling frequency into the precision of volatility estimates and can materially overstate standard errors for Sharpe ratios.

www.ludgerhentschel.com   ludger@ludgerhentschel.com

Contents

Acknowledgements

For helpful comments, I am grateful to John Y. Campbell and Alex Wu. I thank Ingemar Hentschel for valuable research assistance.

1 Introduction

The Sharpe ratio is the dominant summary statistic for risk-adjusted investment performance. By scaling excess returns by volatility, it removes arbitrary leverage choices and places strategies with different risk levels on a common scale. Because investors use Sharpe ratios to allocate capital, evaluate managers, and refine strategies, they care not only about the level of a Sharpe ratio but also about its statistical reliability.

In this paper, we revisit statistical inference for Sharpe ratios using a simple organizing principle: the Sharpe ratio equals the ratio of annualized average returns and annualized volatility, and we learn these components at different statistical rates. Merton (1980) shows that the annualized average return becomes more precise only as calendar time passes, regardless of sampling frequency, whereas annualized volatility becomes more precise as we observe more return realizations. When we observe returns monthly or daily, we estimate volatility from many observations per year. Once we make this distinction explicit, Sharpe ratio inference becomes more transparent and can exploit the additional information about risk contained in higher-frequency data, especially for higher Sharpe ratios.

We organize Sharpe ratio inference around three practical questions. First, does a strategy deliver positive risk-adjusted performance? Second, does one strategy improve upon another? Third, does a strategy exceed a given Sharpe ratio benchmark? Treating these questions separately clarifies both the relevant estimands and the correct standard errors.

For the existence question, a zero Sharpe ratio equals a zero expected excess return. Volatility affects scale but not the logic of the null hypothesis. We therefore test whether the mean return equals zero. Volatility is a nuisance parameter. Testing the ratio rather than the mean introduces estimation noise from risk that plays no role under the null. Direct inference for the mean, using conventional or heteroskedasticity- and autocorrelation-consistent (HAC) standard errors, provides a simpler and more powerful solution. This setting requires no Sharpe-specific variance formulas.

When we compare two positive Sharpe ratios, volatility becomes economically relevant, and we must handle it appropriately. We treat Sharpe ratios as differentiable functions of annualized means and annualized variances and apply a standard delta-method expansion. When we implement this procedure directly from returns, it automatically accommodates heteroskedasticity, serial dependence, and – importantly – the higher statistical precision of volatility estimates when sampling frequency is high. In this framework, Sharpe ratio inference reflects the additional information contained in higher-frequency returns.

Common analytical formulas used in practice rely on iid normal assumptions and express standard errors solely as functions of calendar time T. Sampling frequency does not appear in these formulas. As a result, the underlying statistical machinery cannot incorporate the additional information about risk contained in monthly or daily returns. Implicitly, these formulas treat volatility as if we learned it at the same calendar-time rate as the mean.

In many empirical applications, we estimate volatility from dozens or hundreds of return observations per year. Once we introduce sampling frequency explicitly into the variance expressions, standard errors adjust mechanically to reflect the greater precision of volatility estimates. This additional "dial" materially affects inference when Sharpe ratios are moderate or large.

When we observe underlying returns, we estimate the moment covariance matrix using standard HAC tools, which automatically incorporate the different learning rates of means and variances. When we observe only Sharpe ratios, sampling frequency, and sample span, we provide analytical formulas that explicitly incorporate n. These updated analytical solutions also sharpen intuition about Sharpe ratio tests.

Our contribution is conceptual rather than technical. We do not introduce new asymptotic machinery. Instead, we clarify the object of inference and make the distinction between calendar time and sampling frequency explicit. The existing Sharpe ratio testing literature, beginning with Jobson and Korkie (1981) and refined by Memmel (2003), Lo (2002), and Opdyke (2007), treats the Sharpe ratio itself as the primary object of asymptotic inference. These procedures are asymptotically valid and widely used in practice. We build on the same moment-based foundations but make two distinctions explicit: first, that a zero Sharpe ratio is a mean-zero hypothesis; and second, that volatility typically converges at a different statistical rate than the mean. Making these distinctions explicit yields inference that is more transparent, better aligned with economic interpretation, and more responsive to the information contained in higher-frequency data.

The remainder of the paper develops this framework. Section 2 discusses annualization and establishes notation. Section 3 shows that testing a zero Sharpe ratio reduces to standard inference for the mean. Section 4 develops frequency-aware comparisons of Sharpe ratios using moment-based delta-method inference, contrasts these results with conventional analytical formulas, and specializes the approach to tests against fixed Sharpe thresholds. Section 5 examines the probability that a strategy with positive true Sharpe ratio realizes a low or negative sample Sharpe ratio over finite horizons. Appendix A describes the frequency-aware delta-method variance calculations and derives the corresponding asymptotic analytical standard errors.

2 Sharpe Ratio Estimates

We observe excess returns r_t at frequency n observations per year over T calendar years, so the total number of observations is $N = nT$. Per-period sample moments are

$$\bar{r} = \frac{1}{N}\sum_{t=1}^{N} r_t, \qquad s^2 = \frac{1}{N}\sum_{t=1}^{N}(r_t - \bar{r})^2. \tag{1}$$

Under the usual iid annualization convention, we define the annualized mean, variance, and Sharpe ratio following Sharpe (1994)

$$\mu = n\,\bar{r}, \qquad \sigma^2 = n\,s^2, \qquad S = \frac{\mu}{\sigma}. \tag{2}$$

Throughout, S denotes the annualized Sharpe ratio computed from a finite sample.[1] We denote true or population quantities with an asterisk, when needed.

We begin with these familiar definitions because they are widely used and reported. For liquid assets, such as listed equities or futures, returns typically exhibit little autocorrelation, and these calculations are generally reliable.

When returns are materially autocorrelated, however, the usual variance estimate understates long-run risk. As Lo (2002) shows, positive autocorrelation mechanically inflates Sharpe ratios when we ignore this effect. This issue matters most for illiquid investments, such as certain hedge fund strategies or private assets. In those cases, investors should estimate long-run variance using appropriate methods and compute Sharpe ratios based on the corresponding long-run risk.

We proceed in two steps. First, we develop inference for the standard Sharpe ratio defined in equation (2). This definition is most appropriate for iid returns but is commonly applied outside that setting. If returns are autocorrelated or heteroskedastic, our moment-based inference remains valid for this Sharpe ratio because we estimate the covariance of the underlying

[1] For active investment strategies, Sharpe (1994) shows that the same ratio applied to benchmark-relative (active) returns is often called the information ratio. All statistical results for Sharpe ratios therefore also apply to information ratios.

sample moments using HAC methods. We assume iid returns only when presenting closed-form benchmark formulas for intuition.

Second, if an investor computes Sharpe ratios using HAC long-run risk estimates, the same logic continues to apply but the Sharpe ratio becomes a smooth function of a slightly different set of sample statistics. We estimate the long-run covariance of those statistics and apply the same delta-method machinery. The algebra becomes less compact, but the procedure is identical. We develop this extension in appendix A; the main ideas remain unchanged.

How we define the Sharpe ratio and how we estimate its sampling uncertainty are separate decisions. The analytical formulas in this paper assume iid normal returns for clarity. The inference framework itself does not.

3 Testing the Zero Sharpe Ratio Hypothesis

The most basic and most common Sharpe ratio question is whether performance is positive at all. The null hypotheses of a zero Sharpe ratio and a zero mean are logically equivalent

$$H_0 : S^* = 0 \iff H_0 : \mu^* = 0 \tag{3}$$

because $S = \mu/\sigma$ and volatility satisfies $0 < \sigma < \infty$. Volatility affects the scale of the Sharpe ratio but not the logic of whether it is zero. Under the null, the economically relevant parameter is the expected excess return. Volatility is not part of the hypothesis; it is a nuisance parameter. Testing $S^* = 0$ is therefore a mean-zero test. This equivalence is well known in principle but often overlooked in practice. In academic work, researchers almost always test whether expected returns differ from zero. In practice, however, inference is often framed in terms of Sharpe ratios and implemented using Sharpe-specific standard errors. When framed this way, volatility re-enters the test statistic even though it plays no role in the null hypothesis. Making the equivalence explicit removes unnecessary estimation noise and clarifies the object of inference.

Under standard regularity conditions,

$$\sqrt{T}\,(\mu - \mu^*) \xrightarrow{d} \mathcal{N}(0, \sigma^{*2}), \tag{4}$$

and the conventional test statistic for testing whether the true mean μ^* is zero is

$$Z_\mu = \frac{\mu}{\mathrm{SE}(\mu)}. \tag{5}$$

For iid returns,

$$\mathrm{SE}(\mu) = \frac{\sigma}{\sqrt{T}}, \tag{6}$$

implying

$$Z_\mu = \frac{\mu}{\sigma/\sqrt{T}} = S\sqrt{T}. \tag{7}$$

This is a conventional test for a zero mean that we can express in terms of the Sharpe ratio.

Three features are immediate. First, the effective sample size is calendar time T, not the total number of observations $N = nT$. As Merton (1980) emphasizes, increasing the sampling frequency within a fixed horizon does not improve precision for the annualized mean. Second, the Sharpe ratio enters only as a rescaled mean statistic. No Sharpe-specific variance formula is required. Third, inference requires only the observed Sharpe ratio S and the sample length T.

The existing Sharpe testing literature derives asymptotic standard errors for S itself and does not highlight the test of a zero Sharpe ratio as a special case. Under iid normal returns, the conventional standard errors are

$$\mathrm{SE}(S) = \sqrt{\frac{1 + \frac{1}{2}S^2}{T}}. \tag{8}$$

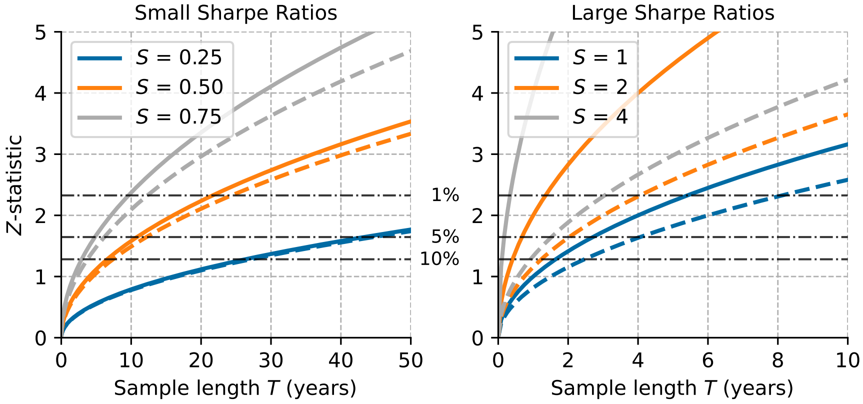
The associated test statistic is

$$Z_S = \frac{S}{\mathrm{SE}(S)} = S\,\frac{\sqrt{T}}{\sqrt{1 + \frac{1}{2}S^2}} = Z_\mu\,\frac{1}{\sqrt{1 + \frac{1}{2}S^2}}. \tag{9}$$

Conventional implementations use these formulas in plug-in form, evaluating the variance at the estimated Sharpe ratio, S. This formula evaluates the variance at the estimated Sharpe ratio, introducing sampling variation in the standard error that is unrelated to the null hypothesis.

Asymptotically, the Sharpe-based test and the mean test coincide under the null hypothesis. When $S^* = 0$, the plug-in variance expression converges to $1/T$ because the estimated Sharpe ratio converges to zero. In this case, Z_S and Z_μ have the same asymptotic distribution. Under the alternative, however, the two statistics do not coincide. If $S^* \neq 0$, the plug-in variance converges to $(1 + S^{*2}/2)/T$, so Z_S differs from Z_μ by a constant scaling factor

Figure 1: Sharpe Ratio Tests for Positivity



The figure compares two asymptotic Z statistics for testing the null hypothesis of a zero Sharpe ratio. The statistic Z_μ, shown with solid lines, is based on a direct test of the mean return. The statistic Z_S, shown with dashed lines, is based on a test of the Sharpe ratio that accounts for sampling error in both the mean and the volatility. Horizontal dash-dot lines indicate the one-sided 1%, 5%, and 10% critical values of the standard normal distribution.

The statistic Z_μ tests the null hypothesis of a zero mean excess return and avoids nuisance sampling variability from estimating volatility, which is irrelevant under the null. The statistic Z_S tests the Sharpe ratio directly and includes sampling variability from the risk estimate. The test statistic Z_S is asymptotically valid but smaller in magnitude than Z_μ, reflecting additional noise that does not correspond to the economic content of the null hypothesis.

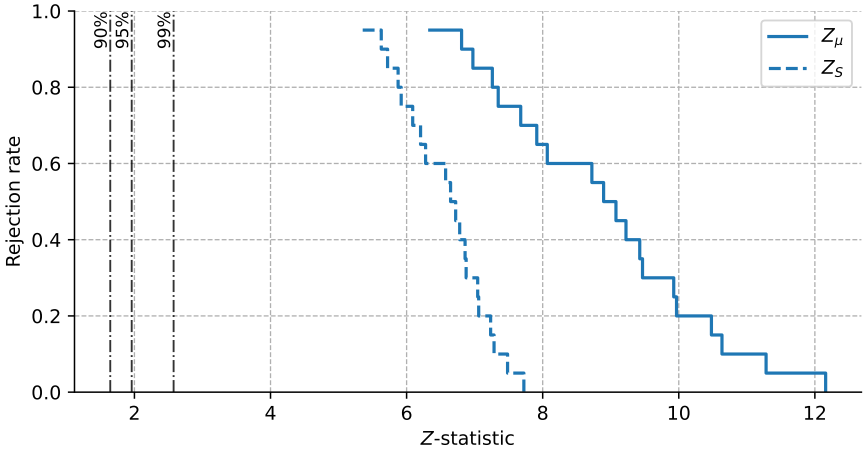
even asymptotically. The difference therefore reflects more than small-sample noise. In finite samples, the plug-in Sharpe test evaluates the variance at the estimated Sharpe ratio rather than under the null. This introduces additional S-dependent scaling that is unrelated to the hypothesis being tested and reduces power relative to the direct mean test.

Figure 1 compares Z_μ and Z_S for different Sharpe ratios and sample sizes. The figure also shows one-sided critical values for reference.[2] At low Sharpe ratios, the statistics behave similarly. At higher Sharpe ratios, which investors consider attractive, Z_μ exceeds Z_S and therefore produces a more powerful test. For example, at a Sharpe ratio of 2, the conventional plug-in test requires roughly four years of iid returns to achieve 1% significance, whereas the mean test reaches the same threshold in about one year. For strategies with high Sharpe ratios, these differences in inference are economically meaningful.

Because Z_μ is the standard test of the mean, its size properties are well understood. The power gain comes from focusing on the parameter that defines the null and not propagating volatility estimation noise that is irrelevant for $H_0 : S^* = 0$.

[2] Economically, we are almost always interested in testing whether the Sharpe ratio is greater than zero. As a result, we use one-sided tests and the associated critical values. A two-sided test may be appropriate if we can short the asset and don't have a meaningful prior expectation about the sign of the average return.

Figure 2: Empirical Sharpe Ratio Tests for Positivity



The figure compares the empirical rejection rates of the mean test Z_μ, shown as a solid blue line, and the conventional plug-in Sharpe ratio test Z_S, shown as a dashed blue line, to test the null hypothesis of a zero Sharpe ratio.

We apply both tests to 20 equally-weighted portfolios, each consisting of 10 non-overlapping signals from the Chen and Zimmermann (2022) signal library. There are 212 signals in the library; we remove the 12 signals with the fewest available observations. In case a return is missing, the portfolios reallocate the weight to the available returns. The returns are monthly from January 1975 to December 2024.

When returns are heteroskedastic or autocorrelated, the same logic applies. We test $\mu = 0$ using a robust estimator of SE(μ), such as the Newey and West (1987) estimator. No Sharpe-specific variance adjustments are required.

Figure 2 provides an empirical illustration. We compare the mean-based statistic Z_μ and the conventional plug-in Sharpe statistic Z_S. We apply both tests to 20 equally-weighted portfolios, each consisting of 10 non-overlapping signals from the Chen and Zimmermann (2022) signal library. There are 212 long-short equity signal returns in the library. After excluding the 12 signals with the fewest overall available observations, each portfolio assigns equal weight to the available returns in its signal set. The data are monthly from January 2015 to December 2024.

Figure 2 plots the empirical rejection rates for both tests. The dashed line shows the conventional plug-in tests; the solid line shows the mean-based test. The mean-based statistic Z_μ is consistently larger than the plug-in Sharpe statistic Z_S. For example, 65% of the portfolios reach or exceed $Z_\mu = 8$, while none reach this level for Z_S. This difference reflects the additional volatility-estimation noise embedded in Z_S.

Because the zero-Sharpe question concerns only the mean, it should be treated separately from a comparison of positive Sharpe ratios. We now turn to that problem.

4 Comparing Sharpe Ratios

Investors frequently compare two strategies with positive Sharpe ratios S_1 and S_2 and ask whether one is statistically larger than the other. Unlike the zero-Sharpe question, this is genuinely a ratio comparison problem and both average return and volatility matter for the ranking. We intentionally do not compare mean excess returns because they may reflect different and potentially arbitrary leverage choices, which we may not observe.

4.1 Looking through the Sharpe ratio

A Sharpe ratio is not a primitive object. It is a differentiable function of two moments, $S = \mu/\sigma$. When we compare S_1 and S_2, we are comparing two functions of four underlying parameters,

$$\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2). \tag{10}$$

The key insight we apply to these estimates is that expected returns and volatilities are learned at different statistical rates. Information about expected returns accumulates with calendar time T. Information about volatility accumulates with the number of return observations $N = nT$. When sampling frequency n is moderate or high, volatility is estimated much more precisely than the mean.

Earlier Sharpe ratio tests treat the Sharpe ratio itself as the object of inference and derive analytical standard errors that scale only with calendar time T. The corresponding formulas leave no role for sampling frequency and therefore omit the additional precision available in volatility estimation.

4.2 Testing Sharpe ratio differences

The conventional comparison is based on the difference in Sharpe ratios[3]

$$D = S_1 - S_2. \tag{11}$$

[3] Economically, common improvements in Sharpe ratios are proportional rather than additive. This suggests comparing $\log S_1 - \log S_2 = (\log \mu_1) - \log \mu_2) - \frac{1}{2}(\log \sigma_1^2 - \log \sigma_2^2)$ and further highlights the separate roles of means and variances. A regularized transformation such as $\frac{1}{2}\log(1 + S^2)$ avoids instability near zero and is a core ingredient in mean-variance efficiency tests like Gibbons, Ross, and Shanken (1989). This leads to frequency-aware variance expressions analogous to the ones we derive here. In practice, once sampling frequency is properly incorporated, proportional and additive tests lead to similar conclusions. For this reason, and to maintain focus on correcting conventional inference, we emphasize the conventional additive comparisons.

Under standard regularity conditions,

$$\sqrt{T}\,(D - D^{*}) \xrightarrow{d} \mathcal{N}(0,\, v_n^2), \tag{12}$$

where v^2 depends on the joint distribution of the underlying moments.

Appendix A extends the delta method approach discussed in Lo (2002) from one to two Sharpe ratios and explicitly includes the effects of sampling frequency. We apply the delta method to the four-dimensional moment vector θ to obtain a variance that separates mean uncertainty from volatility uncertainty in a transparent way. Under iid normal returns sampled at frequency n observations per year over T years, the variance admits a simple closed form

$$Tv_n^2 = 2(1-\rho) + \frac{1}{2n}\left(S_1^2 + S_2^2 - 2\rho^2 S_1 S_2\right), \tag{13}$$

where ρ denotes the contemporaneous correlation between strategy returns.

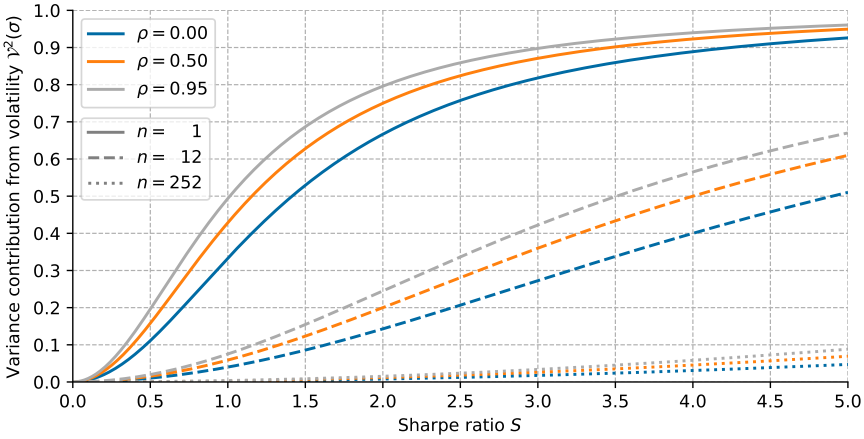
This variance reduces to the solution in Opdyke (2007) when we ignore the frequency dependence of the standard error and set $n = 1$. Our enhancement is to make explicit that volatility is learned at rate nT, not T, and to allow the standard error to respond to that difference in precision.

The structure of this expression is economically intuitive. The first term, $2(1-\rho)$, reflects uncertainty in expected returns.[4] It depends only on calendar time T and is unaffected by sampling frequency. The second term, $\left(S_1^2 + S_2^2 - 2\rho^2 S_1 S_2\right)/(2n)$, reflects volatility estimation error from covariance terms. This component shrinks with sampling frequency n. When returns are observed monthly ($n = 12$) or daily ($n = 252$), volatility uncertainty can be substantially smaller than mean uncertainty.[5]

[4] It is helpful to rewrite the problem as a textbook test for the difference in means with correlated data. Define the difference between volatility-standardized excess returns as $x_t \equiv r_{1,t}/\sigma_1^* - r_{2,t}/\sigma_2^*$, where we treat the true standard deviations σ_k^* as known constants. Then the annualized mean of x_t equals $D = \mu_1/\sigma_1^* - \mu_2/\sigma_2^*$. Under iid sampling, the variance of x_t is $\mathrm{Var}(x_t) = 2(1-\rho)$ because $\mathrm{Var}(r_{k,t}/\sigma_k^*) = 1$, where ρ is the contemporaneous correlation between the returns. In this benchmark case, testing $S_1 = S_2$ is the standard test of whether the mean of x_t equals zero, with test statistic $Z = \sqrt{T}\,D/\sqrt{2(1-\rho)}$.

[5] The framework does not require equal sampling frequencies. If one strategy is observed daily and another monthly, inference should reflect the greater precision of the daily volatility estimate rather than discarding that information. Under the iid normal benchmark with unequal sampling frequencies n_1 and n_2, the variance becomes $Tv^2 = 2(1-\rho) + (S_1^2/n_1 + S_2^2/n_2 - 2\rho^2 S_1 S_2/\sqrt{n_1 n_2})/2$. The volatility component therefore shrinks with each sampling frequency individually, while the covariance term scales with the geometric mean $\sqrt{n_1 n_2}$.

Figure 3: Variance Share of Risk Estimates



The figure plots the variance share attributable to the risk estimate, $\mathcal{V}^2(\sigma)$, in the total asymptotic variance of the difference between two Sharpe ratios. For clarity, we evaluate the expression at $S = S_1 = S_2$. The variance share is independent of sample length T.

The lines correspond to different values of the return correlation ρ and the sampling frequency n, measured in observations per year.

The associated test statistic is

$$Z_D^{(n)} = \frac{\sqrt{T}\,(S_1 - S_2)}{v_n}. \tag{14}$$

4.3 How much uncertainty comes from risk estimation?

The closed-form variance expression in equation (13) separates uncertainty coming from expected returns and uncertainty coming from volatility. To quantify their relative importance, define the risk share of the total Sharpe-ratio variance as

$$\mathcal{V}^2(\sigma) = \frac{\left(S_1^2 + S_2^2 - 2\rho^2 S_1 S_2\right)/(2n)}{2(1-\rho) + \left(S_1^2 + S_2^2 - 2\rho^2 S_1 S_2\right)/(2n)}. \tag{15}$$

The mean share $\mathcal{V}^2(\mu)$ equals one minus this quantity. Importantly, this decomposition does not depend on the sample length T.

For comparisons near the null, where $S_1 \approx S_2 \equiv S$, the expression simplifies. Using $1 - \rho^2 = (1-\rho)(1+\rho)$,

$$\mathcal{V}^2(\sigma) = \frac{S^2(1+\rho)/n}{2 + S^2(1+\rho)/n} = \frac{S^2}{S^2 + 2n/(1+\rho)}. \tag{16}$$

Figure 3 illustrates this variance share of risk estimation as a function of S for several values of ρ and n. Colors encode the return correlation ρ, which has only a modest effect on the variance shares. Line styles – solid, dashed, and dotted – encode the sampling frequency n, which has a dominant effect.

For annual return data ($n = 1$), risk estimates account for roughly half of the variance of Sharpe ratio estimates near $S = 1$. For higher Sharpe ratios, precisely the settings in which investors care most about performance differences, the risk component exceeds 50% and becomes the primary source of estimation noise when $n = 1$.

For daily data ($n = 252$), risk estimates account for less than 10% of total variance for Sharpe ratios below 5. When standard errors incorporate sampling frequency, Sharpe ratio comparisons are then numerically close to tests of differences in means for volatility-standardized returns, treating the standard deviations as effectively known.

Figure 3 decomposes the frequency-aware variance of the Sharpe ratio difference in equation (13) into mean and volatility components. The figure clarifies how sampling frequency n governs the contribution of volatility estimation to total uncertainty. An equivalent interpretation is that procedures that implicitly fix $n = 1$ ignore the additional precision in volatility estimates available at higher sampling frequencies. The vertical distance between the solid and dashed or dotted lines measures how much standard errors tighten once sampling frequency is incorporated.

Near the null, and using frequency-aware standard errors, volatility estimation affects uncertainty only through the ratio

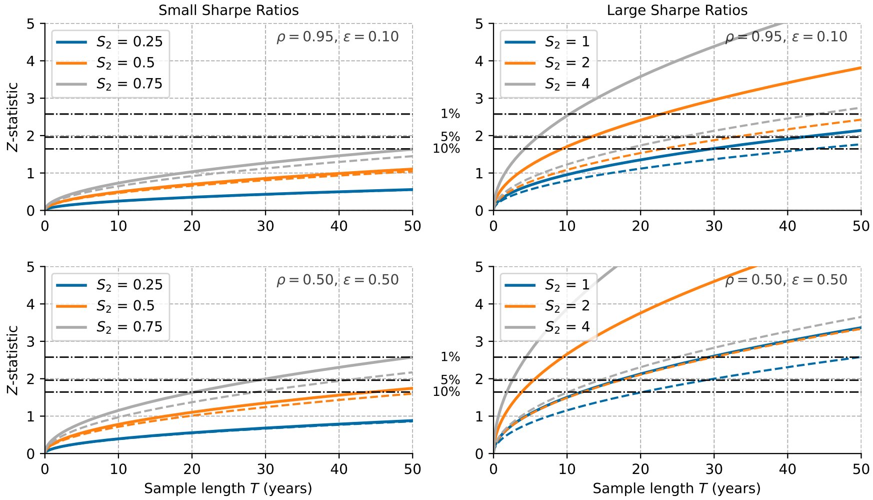
$$q = \frac{S^2(1+\rho)}{n}. \tag{17}$$

When this ratio is small, volatility contributes little to total uncertainty and mean uncertainty dominates the sampling variation of the Sharpe ratio difference. For $0 \leq \rho \leq 1$, the ratio is small when $S^2 \ll n$. In that case, risk estimation plays only a minor role in Sharpe ratio inference.

In typical empirical settings, this condition holds. With monthly data ($n = 12$), $S = 1$, and $\rho = 0.5$, the risk component accounts for less than 6% of total variance. With daily data ($n = 252$), the risk component is well below 1%. Even for moderately high Sharpe ratios, volatility uncertainty is often small relative to mean uncertainty.

Once $S^2 \ll n$, the volatility component of Sharpe ratio variance is negligible relative to mean uncertainty. As we have shown, this condition holds for most empirically relevant Sharpe ratios with daily data ($n = 252$). Pushing to intraday data yields rapidly diminishing statistical benefits but potentially introduces additional complications, like bid-ask bounce or other microstruc-

Figure 4: Tests for Sharpe Ratio Differences: Monthly Data



The figure compares two analytic asymptotic Z statistics for testing the null hypothesis of a zero difference between two Sharpe ratios. The figure illustrates results for a range of Sharpe ratio estimates $S_1 = (1 + \epsilon)S_2$ and assumes monthly return data ($n = 12$) with pairwise correlation ρ.

The statistic $Z_D^{(1)}$, shown with dashed lines, is based on the conventional analytical standard errors for the difference of Sharpe ratios. The statistic $Z_D^{(n)}$, shown with solid lines, is based on analytical standard errors that recognize the sampling frequency n.

Horizontal dash-dot lines indicate the one-sided 1%, 5%, and 10% critical values of the standard normal distribution conditional on having selected $S_1 > S_2$.

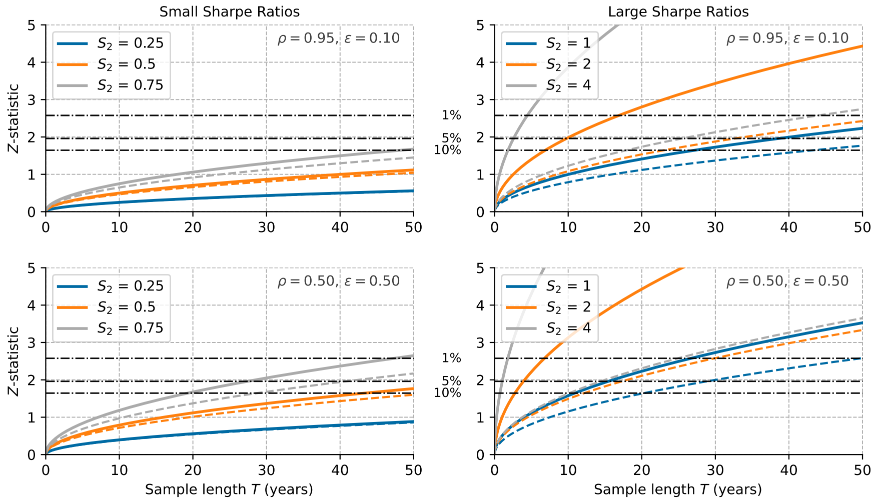
ture noise. In practice, daily data are generally sufficient to make volatility estimation effectively precise for Sharpe ratio inference.

In order for the risk estimates to contribute material uncertainty to the Sharpe ratio standard errors, we need high Sharpe ratios, low sampling frequency, and high correlations. At $S = 1$, $n = 1$, and $\rho = 1$, risk estimates contribute half of the variance and require careful consideration. As soon as we move to monthly returns, however, this variance share drops to 8%.

Sampling frequency governs the precision of volatility estimates and therefore the role of risk uncertainty in Sharpe ratio tests. When we observe monthly or daily returns, we estimate volatility far more precisely than expected return. Whenever higher-frequency data are available, we should use them to estimate risk and compute standard errors that reflect the additional information about risk.

As figure 3 makes clear, sampling frequency determines how much volatility estimation contributes to Sharpe ratio uncertainty. When that contribution is small, n-agnostic standard errors continue to attribute substantial uncertainty to volatility and therefore overstate total variance.

Figure 5: Tests for Sharpe Ratio Differences: Daily Data



The figure compares two analytic asymptotic Z statistics for testing the null hypothesis of a zero difference between two Sharpe ratios. The figure illustrates results for a range of Sharpe ratio estimates $S_1 = (1+\epsilon)S_2$ and assumes daily return data ($n = 252$) with pairwise correlation ρ.

The statistic $Z_D^{(1)}$, shown with dashed lines, is based on the conventional analytical standard errors for the difference of Sharpe ratios. The statistic $Z_D^{(n)}$, shown with solid lines, is based on analytical standard errors that recognize the sampling frequency n.

Horizontal dash-dot lines indicate the one-sided 1%, 5%, and 10% critical values of the standard normal distribution conditional on having selected $S_1 > S_2$.

4.4 Why frequency matters

The variance-share calculations above translate into economically meaningful differences in test statistics.

Figure 4 and figure 5 compare the conventional statistic $Z_D^{(1)}$, shown in dashed lines, to the frequency-aware statistic $Z_D^{(n)}$, shown in solid lines. Both statistics use the analytical standard errors for iid normal returns.

The figures use Sharpe ratio estimates $S_1 = (1+\epsilon)S_2$. Figure 4 uses $n = 12$ for monthly return data and figure 5 uses $n = 252$ for daily return data. In each figure, the left panels show comparisons for smaller Sharpe ratios and the right panels show comparisons for larger Sharpe ratios. In each figure, the top panels show comparisons for highly correlated returns with $\rho = 0.95$ and small proportional gains $\epsilon = 0.1$; the bottom panels show comparisons for less correlated returns with $\rho = 0.50$ and larger proportional gains $\epsilon = 0.5$.[6]

The figures mark critical values for one-sided test conditional on the selection event $S_1 \geq S_2$. Conditioning on $S_1 \geq S_2$ truncates the null distribution to its positive half. A conditional rejection probability of 10% therefore corre-

[6] The tests statistics Z rise with ρ and ϵ. The top and bottom panels combine these effects in offsetting directions.

Table 1: Test Size Simulations

	S_2	$T = 3$	$T = 5$	$T = 10$	$T = 25$	$T = 50$	$T = 100$
Panel A: $\rho = 0.50, n = 12$							
$Z_D^{(1)}$	0.5	0.96	2.60	5.03	7.37	8.01	8.10
	1.0	1.09	2.67	4.10	4.71	4.89	5.00
	3.0	0.03	0.07	0.13	0.17	0.18	0.21
$Z_D^{(n)}$	0.5	6.93	7.37	8.36	9.73	10.00	10.01
	1.0	9.29	9.80	10.22	10.17	10.02	9.95
	3.0	10.77	10.46	10.16	10.09	10.05	9.99
Panel B: $\rho = 0.50, n = 252$							
$Z_D^{(1)}$	0.5	0.60	2.10	4.56	7.12	7.74	7.96
	1.0	0.69	2.11	3.46	4.02	4.30	4.30
	3.0	0.01	0.01	0.01	0.01	0.02	0.01
$Z_D^{(n)}$	0.5	6.17	6.93	8.09	9.64	9.97	10.05
	1.0	8.57	9.39	9.86	9.98	10.02	10.01
	3.0	10.02	9.99	9.99	9.99	9.99	10.05
Panel C: $\rho = 0.95, n = 12$							
$Z_D^{(1)}$	0.5	1.80	3.91	6.03	7.54	8.04	8.24
	1.0	1.32	2.77	4.04	4.78	5.00	5.17
	3.0	0.04	0.11	0.20	0.27	0.31	0.31
$Z_D^{(n)}$	0.5	9.90	9.77	9.87	10.06	9.98	10.06
	1.0	10.46	10.31	10.12	10.04	10.04	9.99
	3.0	10.78	10.34	10.17	10.05	10.05	10.08
Panel D: $\rho = 0.95, n = 252$							
$Z_D^{(1)}$	0.5	1.19	3.11	5.45	7.16	7.67	7.91
	1.0	0.79	2.10	3.31	3.97	4.22	4.33
	3.0	0.00	0.01	0.01	0.01	0.01	0.01
$Z_D^{(n)}$	0.5	9.19	9.31	9.70	9.94	9.94	10.07
	1.0	9.76	9.98	9.96	10.07	9.96	10.03
	3.0	10.10	10.12	9.99	10.06	9.92	9.98

The table reports Monte Carlo simulations with sample size 1 million for Sharpe ratio comparison tests under the null hypothesis that the Sharpe ratios are equal, $S_1 = S_2$. The entries show rejection frequencies at the 10% critical value. For accurately sized tests, we expect to see 10%, apart from simulation noise.

Aligned with common practice, the tests select and test only cases where $S_1 > S_2 > 0$ and condition the one-sided tail probability on these selection events.

The test $Z_D^{(1)}$ tests the difference in Sharpe ratios but is not aware of the sampling frequency n. The test $Z_D^{(n)}$ tests the difference in Sharpe ratios and uses standard errors that are aware of the sampling frequency n. All tests use analytical standard errors for iid returns.

sponds to the 5% upper tail of the unconditional standard normal distribution. This raises the critical values above their conventional, unconditional levels.[7]

When Sharpe ratios are small, both statistics behave similarly because

[7] Our conditioning in these examples addresses the selection event $S_1 \geq S_2$ for a given pair of strategies. We do not consider broader multiple-testing adjustments across many Sharpe ratio

volatility contributes little to total uncertainty. As Sharpe ratios rise, exactly the situations in which investors seek to distinguish strong strategies from merely good ones, $Z_D^{(1)}$ becomes increasingly conservative, especially at low sampling frequencies. With daily data, where volatility uncertainty is negligible relative to mean uncertainty, the two statistics diverge sharply. For example, for Sharpe ratios $S_1 = 3$ and $S_2 = 2$, it takes 30 years of daily returns before $Z_D^{(1)}$ indicates statistical significance at 1%. By contrast, $Z_D^{(n)}$ reaches the same significance level in about one quarter of the time.

Monte Carlo simulations in table 1 confirm that the higher Z values for $Z_D^{(n)}$ shown in the figures reflect improved calibration rather than size distortion. We simulate iid normal returns under the null hypothesis $S_1 = S_2$ and report rejection rates at a 10% critical value. The test statistics use the analytical standard errors shown above. As in the figures, all tests are one-sided and conditioned on the selection $S_1 \geq S_2$, with appropriately adjusted critical values. Each reported rejection rate is based on 1 million simulations.

A correctly sized test should reject approximately 10% of the time. In small samples, particularly for $T < 10$ years, the conventional statistic $Z_D^{(1)}$ rejects far less frequently than 10%. The distortion becomes more severe as Sharpe ratios increase. Even with 100 years of data, $Z_D^{(1)}$ does not fully attain its asymptotic size in the table.

A subtle pattern in table 1 shows that $Z_D^{(1)}$ has slightly different rejection rates for $n = 12$ and $n = 252$, even though its analytical standard error does not depend on n. Under the simulated iid Gaussian returns, the sampling variance of the Sharpe difference $S_1 - S_2$ depends on n because part of its noise arises from estimating risk, whose precision grows with the number of observations $N = nT$. The conventional $Z_D^{(1)}$ standardizes using an n-agnostic approximation and therefore does not allow the volatility-estimation component to shrink with n. As a result, it becomes more conservative as sampling frequency increases. By contrast, the frequency-aware statistic $Z_D^{(n)}$ includes the $1/n$ term and exhibits rejection rates that are stable across sampling frequencies, up to Monte Carlo error.

By contrast, the frequency-aware statistic $Z_D^{(n)}$ delivers rejection rates much closer to the nominal 10% level across sample lengths and Sharpe ratios. The larger Z values in the figures therefore reflect improved statistical efficiency and correct calibration, not over-rejection under the null.

comparisons. Standard procedures, such as false discovery rate control following Benjamini and Hochberg (1995), can be applied directly to the critical values or the resulting p-values.

4.5 Implementation with general returns

The analytical formulas above assume iid normal returns and serve as useful benchmarks. This assumption combines two distinct restrictions. The iid assumption rules out heteroskedasticity and serial dependence in returns, while the normality assumption rules out fat tails and other higher-moment deviations from the Gaussian distribution. In practice, returns may violate both restrictions. Returns may not be iid because they are heteroskedastic or autocorrelated. Returns may also not be normally distributed, possibly because they contain jump components that induce fat tails.

When underlying returns are available, we follow Lo (2002) and estimate the covariance matrix of the sample moments directly and apply the delta method. The difference in Sharpe ratios is a differentiable function of annualized means and variances. We compute its asymptotic variance using the covariance matrix of those moments.

This moment-based approach requires no Sharpe-specific variance formulas and does not rely on iid or normality assumptions. The appendix provides the explicit derivations and shows how sampling frequency enters the variance through the moment covariance matrix.

These calculations accommodate a range of deviations from iid normal returns. In particular, covariance estimators such as Newey and West (1987) handle autocorrelation and heteroskedasticity in returns. The moment covariance matrix also incorporates higher moments that arise when returns are not normally distributed, for example when returns contain jumps in addition to normally distributed diffusion components.

Return jumps affect the precision of estimates for both mean returns and risk. In settings where such jumps are economically important, such as strategies that sell deep out-of-the-money options, Goetzmann, Ingersoll, Spiegel, and Welch (2007) and Martin and Shi (2026) show that empirical Sharpe ratios can be fragile performance measures. Many investors recognize this and interpret Sharpe ratios with extreme care in these situations.

Conceptually, jumps in returns prevent us from realizing the full potential asymptotic benefit of higher sampling frequencies. Jumps induce higher moments that we do not learn at the same rate as return risk. In effect, these higher-moment contributions create a precision ceiling that cannot be breached simply by increasing the sampling frequency. The importance of these effects depends on the size and frequency of the jumps. Fortunately, the delta-method standard errors in the appendix incorporate these higher-moment effects and produce larger standard errors than the analytical benchmarks when this is appropriate.

In the main text we focus on cases where near-diffusion returns dominate, Sharpe ratios remain appropriate summary measures of performance, and daily return data allow us to come close to the high-frequency asymptotic limit.

4.6 Testing against a positive Sharpe threshold

Testing whether a strategy exceeds a benchmark Sharpe ratio $H_0 : S^* = S_0 > 0$ is a special case of the comparison problem. Setting $S_2 = S_0$ and treating it as fixed yields

$$T\,\mathrm{Var}(S - S_0) = 1 + \frac{1}{2n}\,S^2. \tag{18}$$

If we ignore the sampling frequency by setting $n = 1$, this reduces to the analytical variance for iid normal returns derived in Lo (2002).

For threshold tests, we can evaluate the standard error either at the estimated Sharpe ratio S (Wald form) or under the null hypothesis S_0 (score form). Evaluating the standard error under the null replaces S^2 by S_0^2 and yields

$$Z_n(S_0) = \frac{\sqrt{T}\,(S - S_0)}{\sqrt{1 + \frac{1}{2n} S_0^2}}. \tag{19}$$

Once again, the frequency term reflects the precision with which volatility is estimated. When n is large, this correction can materially tighten inference relative to standard Sharpe-based formulas.

Evaluating the standard error under the null hypothesis is particularly attractive when S_0 is small. As S_0 approaches zero, the null-imposed statistic is continuous with the mean test because the volatility-estimation correction vanishes and equation (19) reduces to the mean test in equation (7). In contrast, evaluating the standard error at S introduces S-dependent scaling even near the boundary, where volatility uncertainty plays little role in the hypothesis. In practice, the null-imposed version also avoids having to switch test types at $S_0 = 0$. Asymptotically, the Wald and score forms are equivalent.[8]

Table 2 compares these standard errors for a range of sampling frequencies n, sample periods T, and Sharpe ratios S_0. For $n = 1$, these standard errors match the values shown in Table 1 of Lo (2002). The table confirms that

[8] For tests of $S_1^* = S_2^*$, evaluating standard errors under the null would require a restricted estimator that enforces $\mu_1/\sigma_1 = \mu_2/\sigma_2$. Unlike threshold tests, this restriction does not yield a single canonical plug-in choice without additional structure (for example, a likelihood specification or an explicit weighting of the underlying moment conditions).

Table 2: Standard Errors for Threshold Sharpe Tests

S_0	n	$T = 3$	$T = 5$	$T = 10$	$T = 25$	$T = 50$	$T = 100$
0.5	1	0.612	0.474	0.335	0.212	0.150	0.106
	12	0.580	0.450	0.318	0.201	0.142	0.101
	252	0.577	0.447	0.316	0.200	0.141	0.100
	∞	0.577	0.447	0.316	0.200	0.141	0.100
1.0	1	0.707	0.548	0.387	0.245	0.173	0.122
	12	0.589	0.456	0.323	0.204	0.144	0.102
	252	0.578	0.448	0.317	0.200	0.142	0.100
	∞	0.577	0.447	0.316	0.200	0.141	0.100
1.5	1	0.842	0.652	0.461	0.292	0.206	0.146
	12	0.604	0.468	0.331	0.209	0.148	0.105
	252	0.579	0.448	0.317	0.200	0.142	0.100
	∞	0.577	0.447	0.316	0.200	0.141	0.100
2.0	1	1.000	0.775	0.548	0.346	0.245	0.173
	12	0.624	0.483	0.342	0.216	0.153	0.108
	252	0.580	0.449	0.317	0.201	0.142	0.100
	∞	0.577	0.447	0.316	0.200	0.141	0.100

The table reports standard errors for the test that the true Sharpe ratio S^* equals a known value S_0, $H_0 : S^* = S_0$. For iid normal returns, the asymptotic standard error under the null hypothesis is $([1 + S_0^2/(2n)]/T)^{1/2}$, where T is the sample length in calendar years and n counts observations per year. When $n = 1$, the standard error expression reduces to the conventional calendar-only form reported in Lo (2002).

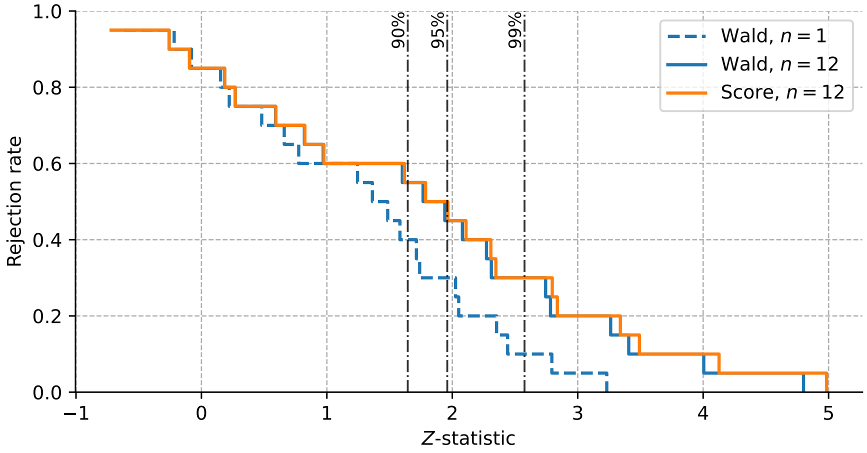
daily data with $n = 252$ generally provide precision that is very close to the limit $n \to \infty$, without introducing complications that may arise with higher-frequency intraday return data, such as bid-ask bounce or other microstructure noise. As n increases, the volatility-estimation correction vanishes and the frequency-aware standard error converges to $1/\sqrt{T}$, which corresponds to testing the mean of volatility-standardized returns when the standard deviation is effectively known.

4.7 Empirical illustration

Figure 6 compares empirical rejection rates for the frequency-aware threshold test and the conventional plug-in Sharpe ratio tests of Lo (2002) and Opdyke (2007). We test the same 20 portfolios we used in figure 2. Each portfolio consists of 10 single-signal portfolios drawn from the 212 long-short equity anomaly portfolios in Chen and Zimmermann (2022) and the returns are monthly from January 1975 to December 2024.

Figure 6 tests the portfolio Sharpe ratios against the threshold $S_0 = 1$. The dashed blue line shows the conventional plug-in test assuming $n = 1$, denoted $Z_1(S_0)$, with standard errors evaluated at the estimated Sharpe ratios (Wald form). The solid lines show the frequency-aware statistic $Z_n(S_0)$ using $n = 12$,

Figure 6: Empirical Sharpe Ratio Tests for $S_0 = 1$



The figure compares the empirical rejection rates of the frequency-aware Sharpe ratio test $Z_n(S_0 = 1.0)$, using solid lines, and the conventional Sharpe ratio test $Z_1(S_0 = 1.0)$, using a dashed line. For the frequency-aware test, the figure shows results for the Wald form in bue, with standard errors evaluated at the Sharpe ratio estimates S, and the score form in orange, with standard errors evaluated at the null hypothesis S_0. All tests use asymptotic analytical standard errors assuming iid normal returns.

We apply both tests to 20 equally-weighted portfolios, each consisting of 10 non-overlapping signals from the Chen and Zimmermann (2022) signal library. There are 212 signals in the library; we remove the 12 signals with the fewest available observations. In case a return is missing, the portfolios reallocate the weight to the available returns. The returns are monthly from January 1975 to December 2024.

corresponding to the monthly return data. The blue solid line reports the Wald form and the orange solid line reports the score form, which evaluates standard errors under the null hypothesis.

The dominant effect comes from incorporating sampling frequency. Moving from $n = 1$ to $n = 12$ materially increases empirical rejection rates. For example, at the 95% confidence level, the frequency-aware test rejects 50% of the hypotheses, compared to 30% for the conventional plug-in test. At the 99% confidence level, empirical rejection rates are 30% versus 10%.

At monthly frequency, volatility contributes only modestly to overall uncertainty. As a result, the Wald and score versions of the frequency-aware test are very similar. This supports using the score form in practice, which avoids special treatment of $S_0 = 0$ and ensures continuity with the mean test at the boundary.

4.8 Bootstrap inference

An alternative approach to inference for Sharpe ratios is to use bootstrap methods (see Efron and Tibshirani (1994) for an overview). By resampling the observed return series and recomputing the statistic, the bootstrap approximates the finite-sample distribution without relying on closed-form asymptotic formulas. For smooth statistics such as Sharpe ratios, bootstrap methods are asymptotically valid under standard regularity conditions.

However, bootstrap procedures inherit the convergence structure of the statistic they approximate. The Sharpe ratio is a smooth function of the mean and the variance, and these components are learned at different statistical rates: the mean converges at rate $T^{-1/2}$, while the volatility estimate converges at rate $(nT)^{-1/2}$. Standard bootstrap implementations resample the underlying observations and therefore replicate the joint sampling variation of all components, but they do not make transparent this separation of rates or how sampling frequency alters the relative precision of volatility.

In particular, the bootstrap does not automatically remove volatility-estimation noise when it is irrelevant to the hypothesis being tested. If one bootstraps a Sharpe-ratio Wald statistic, the resampling procedure reproduces the same volatility-dependent scaling embedded in that statistic. Nor does the bootstrap clarify when volatility uncertainty is economically negligible relative to mean uncertainty.

Under serial dependence, bootstrap inference requires additional tuning choices, such as block-length selection. Block resampling reduces the effective number of independent observations and can materially affect precision when the calendar span T is limited. These choices introduce additional complexity without isolating the distinct roles of mean and volatility uncertainty.

Our results clarify how bootstrap procedures should be structured for Sharpe ratio tests. However, expressing the Sharpe ratio as a smooth function of means and variances yields direct, transparent, and computationally simple standard errors via the delta method. This approach makes explicit how sampling frequency affects inference and accommodates HAC covariance estimation without additional resampling steps.

5 Probability of Falling Below a Sharpe Threshold

Even when a strategy has genuinely positive performance, realized Sharpe ratios can fall below economically relevant hurdle rates over finite samples. Investors often ask a practical question: if the true Sharpe ratio is $S_1 > 0$, what is the probability that a T-year sample produces an estimate below a threshold S_0?

Formally, define the downside probability

$$\pi(S_0; S_1) \equiv \Pr(S \leq S_0 \mid S^* = S_1), \qquad S_1 > 0, \tag{20}$$

where S is the annualized sample Sharpe ratio computed using n observations per year over T calendar years. This is not a hypothesis test. It is a sampling-risk measure that quantifies how often finite samples understate true performance.

5.1 Asymptotic approximation

Under standard regularity conditions,

$$\sqrt{T}\,(S - S^*) \xrightarrow{d} \mathcal{N}(0, v_S^2), \tag{21}$$

where we obtain v_S^2 by applying the delta method to the moment vector (μ, σ^2). A normal approximation yields

$$\pi(S_0; S_1) \approx \Phi\left(\sqrt{T}\,\frac{S_0 - S_1}{v_S}\right), \tag{22}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The probability that S falls below S_0 depends on the distance between S_1 and S_0, scaled by the sampling uncertainty of S, and grows with $\sqrt{T}$.

5.2 Frequency-aware closed form under iid normal returns

Under iid Gaussian returns sampled at frequency n observations per year, the frequency-aware delta-method variance simplifies to

$$v_S^2 = 1 + \frac{1}{2n}\,S_1^2. \tag{23}$$

Substituting into equation (22) gives

$$\pi(S_0; S_1) \approx \Phi\left(\sqrt{T}\,\frac{S_0 - S_1}{\sqrt{1 + \frac{1}{2n}S_1^2}}\right). \tag{24}$$

Two features are immediate. First, mean uncertainty scales with calendar time T. Even with very high sampling frequency, learning about expected returns requires time. Second, volatility uncertainty shrinks with sampling frequency n. When risk is estimated from many observations per year, the term $S_1^2/(2n)$ becomes small and volatility estimation contributes little to overall uncertainty.

Downside probabilities summarize the risk that finite-sample performance appears to fall short of a benchmark even when true performance exceeds that benchmark. They provide a simple and economically transparent tool for performance monitoring, capital allocation decisions, and manager evaluation.

Most importantly, the frequency-aware expression in equation (24) makes clear that uncertainty is typically dominated by mean estimation error, which accumulates only with calendar time, while volatility estimation error rapidly becomes negligible when risk is estimated from high-frequency data.

5.3 Implementation under dependence

In practice, we estimate downside probabilities using the same moment-based covariance framework used for Sharpe ratio tests. We estimate the long-run covariance matrix of the sample mean and variance using either the iid formula or a HAC estimator applied to the centered moment series $\psi_t = \{r_t - \overline{r},\ (r_t - \overline{r})^2 - s^2\}$. We then compute v_S^2 using the analytical gradient of $S = \mu/\sigma$ and evaluate equation (22).

6 Conclusion

This paper clarifies the role of uncertainty from risk estimates in Sharpe ratio tests. The Sharpe ratio combines expected return and volatility, but these components are learned with very different precision. By separating these sources of uncertainty, we can remove volatility noise when it is irrelevant and reduce it when it is unavoidable.

For the most basic question, whether performance is positive, the null hypothesis that the Sharpe ratio equals zero is simply the hypothesis that expected excess return equals zero. Volatility is not part of that hypothesis. Any procedure that treats the Sharpe ratio itself as the object of inference introduces unnecessary estimation noise from the risk estimate. A direct test of the mean eliminates this noise and yields strictly more efficient inference in finite samples.

For comparisons among positive Sharpe ratios and tests against performance thresholds, volatility is economically relevant. Its statistical contribution, however, depends on how precisely it is measured. In common applications, volatility is estimated from many observations per year, making it far more precise than expected return. Conventional analytical formulas do not reflect this difference in precision and therefore tend to overstate standard errors, especially in the range of Sharpe ratios that investors consider economically compelling. We show that with daily data, volatility estimation uncertainty is typically negligible relative to mean uncertainty. In that case, Sharpe ratio comparisons become numerically close to tests for differences in means of volatility-standardized returns, as if the standard deviations were known.

A moment-based delta-method framework makes the separate roles of mean and volatility explicit. When return data are available, standard covariance estimators automatically incorporate heteroskedasticity, autocorrelation, and sampling frequency. When only summary statistics are available, frequency-aware analytical formulas provide transparent and practical alternatives.

Sharpe ratio inference should reflect how precisely risk is measured. When we recognize and manage the contribution of risk-estimation uncertainty, Sharpe ratio tests become more transparent and more accurate.

7 References

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, 2003, Modeling and forecasting realized volatility, *Econometrica* 71 (2), 579–625.

Benjamini, Yoav, and Yosef Hochberg, 1995, Contolling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57 (1), 289–300.

Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review* 11 (02), 207–264.

Efron, Bradley, and Robert J. Tibshirani, 1994, *An Introduction to the Bootstrap* (Chapman and Hall/CRC, New York, NY).

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 57 (5), 1121–1152.

Goetzmann, William, Jonathan Ingersoll, Matthew Spiegel, and Ivo Welch, 2007, Portfolio performance manipulation and manipulation-proof performance measures, *Review of Financial Studies* 20 (5), 1503–1546.

Jobson, J. D., and Bob Korkie, 1981, Performance hypothesis testing with the Sharpe and Treynor measures, *Journal of Finance* 36 (4), 889–908.

Lo, Andrew W., 2002, The statistics of Sharpe ratios, *Financial Analysts Journal* 58 (4), 36–52.

Martin, Ian W., and Ran Shi, 2026, On the moments of the stochastic discount factor, Working paper, London School of Economics, London, England.

Memmel, Christoph, 2003, Performance hypothesis testing with the Sharpe ratio, *Finance Letters* 1 (1), 21–23.

Merton, Robert C., 1980, On estimating the expected return on the market, *Journal of Financial Economics* 8, 323–361.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55 (3), 703–708.

Opdyke, Jeff D., 2007, Comparing Sharpe ratios: So where are the p-values?, *Journal of Asset Management* 8 (5), 308–336.

Sharpe, William F., 1994, The Sharpe ratio, *Journal of Portfolio Management* 21 (1), 49–58.

A Frequency-Aware Standard Errors

This appendix derives a frequency-aware delta-method standard error for differences in Sharpe ratios, $D = S_1 - S_2$, based on the joint asymptotic distribution of annualized means and variances.

We use the delta-method framework described in Lo (2002), which is convenient and directly compatible with standard HAC covariance estimation. We extend this framework to differences in Sharpe ratios. Existing presentations of Sharpe ratio differences are typically given in closed form under iid assumptions (e.g., Opdyke (2007)), without an explicit moment-based expansion in the underlying means and variances. Writing the problem in terms of the joint moment vector makes the role of sampling frequency transparent and allows straightforward incorporation of HAC covariance estimates.

A.1 Annualized inputs

We observe excess returns $r_{k,t}$ for strategies $k \in \{1, 2\}$ at frequency n observations per year over a calendar span of T years, so $N = nT$. We compute per-period sample moments

$$\bar{r}_k = \frac{1}{N}\sum_{t=1}^{N} r_{k,t}, \qquad s_k^2 = \frac{1}{N}\sum_{t=1}^{N}(r_{k,t} - \bar{r}_k)^2, \tag{25}$$

and we annualize under the iid scaling convention,

$$\mu_k = n\,\bar{r}_k, \qquad \sigma_k^2 = n\,s_k^2, \qquad S_k = \frac{\mu_k}{\sigma_k}. \tag{26}$$

We define the Sharpe difference

$$D \equiv S_1 - S_2 = \frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}. \tag{27}$$

The linear annualization $\mu_k = n\bar{r}_k$ and $\sigma_k^2 = ns_k^2$ corresponds to the usual iid scaling convention. The closed-form results below rely on this structure. However, the general moment-based delta-method estimator in (34) does not require iid returns. When returns are heteroskedastic or serially correlated, we estimate the long-run covariance matrix of the sample moments directly using a HAC estimator. In that case, we interpret the annualized variance as the long-run variance of returns, and do not impose an iid scaling assumption on the covariance estimator.

Next, we express the Sharpe-based estimand D as a differentiable function of annualized means and variances. Then, we obtain the joint asymptotic distribution of these moments under calendar-time scaling, making the role of sampling frequency explicit. Finally, we apply the delta method to obtain a feasible standard error.

A.2 Moment vector and covariance estimation

The per-period centered moment vector is,

$$\widetilde{\psi}_t \equiv \begin{pmatrix} r_{1,t} - \bar{r}_1 \\ (r_{1,t} - \bar{r}_1)^2 - s_1^2 \\ r_{2,t} - \bar{r}_2 \\ (r_{2,t} - \bar{r}_2)^2 - s_2^2 \end{pmatrix}, \qquad t = 1, \ldots, N, \tag{28}$$

and we define the corresponding per-period moment vector

$$\widetilde{\theta} \equiv (\bar{r}_1, s_1^2, \bar{r}_2, s_2^2)^\top. \tag{29}$$

We convert to annualized moments via $\theta = A\widetilde{\theta}$ with $A \equiv \mathrm{diag}(n, n, n, n)$, where A reflects the iid annualization convention for both means and variances.

We estimate the long-run covariance of $\widetilde{\theta}$ using the time series $\{\widetilde{\psi}_t\}_{t=1}^N$. In the iid benchmark, we use the sample covariance

$$\widetilde{\Omega} = \frac{1}{N} \sum_{t=1}^{N} \widetilde{\psi}_t \widetilde{\psi}_t^\top. \tag{30}$$

When returns exhibit heteroskedasticity or serial dependence, we replace this with a standard HAC estimator applied to $\widetilde{\psi}_t$ (for example, Newey and West (1987)). We then scale to annualized units,

$$\Omega = n\,\widetilde{\Omega}. \tag{31}$$

Because annualized means and variances equal n times their per-period counterparts, their covariance matrix scales linearly with n.

A.3 Delta-method variance

We write $D = g(\theta)$ with

$$g(\theta) = \frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}. \tag{32}$$

The gradient of g with respect to $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^\top$ has the closed form

$$\nabla g(\theta) = \begin{pmatrix} \partial h/\partial\mu_1 \\ \partial h/\partial\sigma_1^2 \\ \partial h/\partial\mu_2 \\ \partial h/\partial\sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^{-1} \\ -\frac{1}{2}\mu_1\sigma_1^{-3} \\ -\sigma_2^{-1} \\ \frac{1}{2}\mu_2\sigma_2^{-3} \end{pmatrix} = \begin{pmatrix} \sigma_1^{-1} \\ -\frac{1}{2}S_1\,\sigma_1^{-2} \\ -\sigma_2^{-1} \\ \frac{1}{2}S_2\,\sigma_2^{-2} \end{pmatrix}. \tag{33}$$

We compute the feasible variance estimator

$$v_D^2 = \nabla g(\theta)^\top\,\Omega\,\nabla g(\theta), \tag{34}$$

and we form the test statistic

$$Z_D^{(1)} = \frac{D}{v_D}\sqrt{T}. \tag{35}$$

This recipe makes implementation direct: we compute $\widetilde{\psi}_t$, we estimate $\widetilde{\Omega}$ with an iid or HAC tool, we scale via (31), and we evaluate (34) using the analytical gradient (33).

The scaling from $\widetilde{\Omega}$ to $\Omega = n\widetilde{\Omega}$ reflects the change from per-period to annualized moments and does not rely on iid returns. Serial dependence is already incorporated in $\widetilde{\Omega}$ through the HAC estimator.

A.4 Closed-form solution

We provide a closed-form expression for the variance as an analytical benchmark, assuming returns are iid bivariate normal. In applications, we recommend the general moment-based estimator in (34), which remains valid under non-normality, heteroskedasticity, and serial dependence.

Suppose $(r_{1,t}, r_{2,t})$ are iid bivariate normal at the sampling frequency with contemporaneous correlation ρ. Then the annualized covariance matrix of $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^\top$ takes the form

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \rho\sigma_1\sigma_2 & 0 \\ 0 & \frac{2}{n}\sigma_1^4 & 0 & \frac{2}{n}\rho^2\sigma_1^2\sigma_2^2 \\ \rho\sigma_1\sigma_2 & 0 & \sigma_2^2 & 0 \\ 0 & \frac{2}{n}\rho^2\sigma_1^2\sigma_2^2 & 0 & \frac{2}{n}\sigma_2^4 \end{bmatrix}. \tag{36}$$

If returns are not bivariate normal, the mean-variance cross terms don't generally vanish. The delta method based on the analytical gradients and empirical moment matrix accounts for this.

Combining (33) and (36) yields

$$T \ \mathrm{Var}(D) = 2\left(1 - \rho\right) + \frac{1}{2n}\left(S_1^2 + S_2^2 - 2\rho^2 S_1 S_2\right), \tag{37}$$

where S_1 and S_2 denote the annualized sample Sharpe ratios. The first term reflects uncertainty in mean returns, which scales with calendar time. The second term reflects uncertainty in volatility estimates, which shrinks with sampling frequency through the factor $1/n$.

The delta-method approximation for D relies on a first-order linearization in the estimation errors of (μ_k, σ_k). In our setting, this approximation is especially accurate because the mean and volatility enter with very different statistical precision and we account for this.

Information about annualized expected returns accumulates with calendar time T, while information about annualized volatility accumulates with the number of return observations $N = nT$. As a result, volatility estimates become substantially more precise than mean estimates as the sampling frequency increases, even when the calendar span of the data is fixed. This distinction has been recognized at least since Merton (1980) and underlies the use of high-frequency data for risk estimation described in Andersen, Bollerslev, Diebold, and Labys (2003).

Because annualized mean estimates converge at rate $T^{-1/2}$ while annualized volatility estimates converge at rate $(nT)^{-1/2}$, volatility estimation error is smaller by a factor of $n^{-1/2}$. Cross-terms inherit this smaller order. As a result, the dominant component of the variance of D comes from mean uncertainty, with a smaller additive contribution from volatility uncertainty that shrinks with sampling frequency. This structure makes it straightforward to incorporate the additional precision available when returns are observed at higher frequency.

This separation of precision is what makes the delta-method variance for D compact and transparent in the present setting. It also explains why the resulting standard errors differ from the single-frequency closed-form expressions in Lo (2002) and Opdyke (2007), which abstract from sampling frequency and therefore do not distinguish between the learning rates of means and volatilities.